IJASST
International Journal of
Applied Sciences and Smart Technologies
p-ISSN 2655-8564, e-ISSN 2685-943, pp. 43-56, 2026

Volume 08
Issue 01

Handling Highly Imbalanced Flood Data Using K-Means Clustering in Skyline Query Dominance Testing

Vega Purwayoga[1*], Zakwan Gusnadi [1], Winda Ayu Anggraini[2]

[1]*Faculty of Engineering, Siliwangi University, Indonesia*
[2]*Faculty of Economics and Business, Siliwangi University, Indonesia*
[*]*Corresponding Author: vega.purwayoga@unsil.ac.id*

(Received 21-09-2025; Revised 30-10-2025; Accepted 16-04-2026)

Abstract

Skyline query is a recommendation algorithm used to select objects based on multi-attribute preferences, but a key challenge is that its results can be highly imbalanced, where only a small number of objects meet the preferred criteria. This imbalance reduces the reliability of spatial decision-making, including in flood vulnerability assessment. This study addresses the issue by applying a modified Sort-Filter Skyline method that considers maximum and minimum attribute preferences during sorting. The skyline output shows a strong class imbalance, with only 18 areas identified as flood-prone compared to 1,574 non-flood-prone areas. To mitigate this, K-Means clustering is used as a refinement step. The Elbow and Gap Statistic methods recommend three clusters as optimal, while the Silhouette method suggests eight. Cluster distribution analysis shows that three clusters produce a more balanced representation, with Scheme 1 and Scheme 3 showing better balance ratios and lower variation than Scheme 2. Thus, clustering into three groups helps achieve a more representative mapping of flood-prone areas.

Keywords: Flood, Imbalanced Data, K-Means Clustering, Skyline Query

1 Introduction

High rainfall intensity, lowland topography, and interregional river flows are key factors that increase vulnerability to flooding. Therefore, flood potential identification plays a crucial role, not only in mapping high-risk areas but also in providing a foundation for disaster mitigation planning, strengthening community resilience, and setting local intervention priorities. By applying data-driven vulnerability analysis, utilizing both tabular sources and information from online news, flood potential mapping can provide a more comprehensive picture of the risks faced while also contributing to global agendas on disaster risk reduction and sustainable development [1].

The search and recommendation of optimal locations for various case studies can be performed using a skyline query [2], [3]. Skyline query works by examining the dominance of each factor associated with an object or location [4], [5]. In the context of disaster management, particularly floods, Skyline Query can be applied to evaluate locations based on multiple attributes such as elevation, rainfall, land slope, and land cover [6], [7]. In data science, the function of a skyline query is to narrow the search space so that the computational time required to identify relevant objects can be optimized [8], [9].

Skyline Query is effective in selecting the most relevant information, ensuring that the recommended results align with predefined preferences. For instance, in the case of flood disasters, a skyline query may recommend areas with the lowest elevation, the highest rainfall, and/or the flattest slope [7]. This dominance testing results in only a small number of areas being labeled as flood-prone compared to those classified as non-flood-prone [10], [11]. Consequently, the recommendations lead to imbalanced data.

Data imbalance, one of the outcomes produced by a skyline query, presents a significant challenge in classification tasks. Ideally, the knowledge derived from data classification should come from a dataset with balanced labels, ensuring that no particular label disproportionately influences the model [12], [13]. The imbalance resulting from skyline queries can be addressed by applying clustering-based rebalancing, in which data is grouped into several clusters according to the number of clusters determined or recommended [14].

One of the most popular and widely used clustering algorithms is K-Means, which partitions data into groups based on their proximity to cluster centers (centroids). K-Means is advantageous due to its high computational efficiency and its ability to handle large-scale datasets [15], [16]. One approach to addressing data imbalance is by determining the appropriate number of clusters in K-Means. Several well-established methods for recommending the optimal number of clusters (K) include the Elbow, Silhouette, Davies-Bouldin Index, and Calinski-Harabasz methods [17]. Study [18] evealed that the two most effective methods for optimizing K are the Elbow and Silhouette approaches, both of which recommended the same K value in the case of disaster data clustering.

Based on the imbalanced data problem generated by the skyline query algorithm, this study proposes a solution to address the issue. The first objective of this research is to apply the skyline query algorithm to recommend flood-prone areas. The second objective is to implement K-Means optimized using the Elbow and Silhouette methods. The data clustered with K-Means will form several new groups according to the number of clusters determined. The level of data balance will then be measured based on the difference in the number of objects within each cluster.

2 Material and Methods

Studi Area and Research Data

The study area in this research covers several regions in Central Java Province, including Batang, Blora, Boyolali, Demak, Grobogan, Jepara, Karanganyar, Kendal, Klaten, Magelang City, Semarang City, Kudus, Magelang, Pati, Rembang, Semarang, Sragen, Sukoharjo, and Temanggung. These regions are identified as frequently experiencing flood events according to data from National Disaster Management Agency (BNPB). This research consists of several stages: data acquisition, dominance testing using the skyline query algorithm, application of K-Means clustering, and visualization of data distribution.

Data Acquisition

Some of the data acquired in this study include administrative boundaries, elevation, land cover, slope, and rainfall. Administrative boundary data were obtained from the Geospatial Information Agency (BIG) and then filtered to select only a number of sample regions. Elevation and slope data were sourced from National Digital Elevation Model (DEMNAS), which are in raster format and therefore required several preprocessing steps to generate appropriate classes and weights. Rainfall data were collected from multiple weather stations located in Central Java Province, with the values from each station interpolated across the entire study area. Another critical dataset for identifying flood-prone areas is land cover, which helps determine land use within the study area. The smaller the proportion of infiltration areas in a region, the higher its flood risk [19].

Data Integration

In the data integration stage, all datasets were standardized into the same coordinate reference system and clipped using the administrative boundaries as the study area reference. Elevation and slope data from DEMNAS were reclassified to form analysis classes with assigned weights. Rainfall data from multiple weather stations were interpolated to produce a continuous rainfall surface. Land cover data were simplified into broader categories relevant to flood assessment. Finally, all processed layers were combined through spatial overlay to form a single integrated dataset for analysis.

Dominance Testing using Skyline Query

This study applies a modified Sort-Filter Skyline (SFS) algorithm. In terms of performance, SFS is faster in recommending an object as part of the skyline set because it performs sorting in advance [20]. The modification introduced in this research involves adjusting the formula when an attribute preference is set to minimum, allowing the recommendation results to better align with the criteria for flood potential. The modification can be seen in line 2 of Algorithm 1.

Algorithm 1. Res-Q

Input: Dataset D
Output: The Set of skyline points of *D*

```
1:  D ← if ("D[ai]= minimum preference") then
2:          1 – normalize D[ai]
3:      else
4:          normalize D[ai]
5:      end if
6:  E(D) ← calculate entropy of D[ai] using (3)
7:  D ← sort dataset by descending D[ai]
8:  S ← data with the highest entropy D
9:  From 1 to D
10:     if ("D is not dominated") then
11:         write (S, D)
12:     else
13:         remove (S, D)
14:     end if
15: end
```

The modification is applied during the normalization process, as normalization directly affects the entropy calculation. Entropy is one of the values used to determine whether an object is more highly recommended or not. When the attribute preference is set to minimum, smaller values should be considered more preferable; therefore, the normalization formula must be inverted. This adjustment appears in line 2 of Algorithm 1, where minimum preference uses reverse normalization. For example, regions with lower elevation are more likely to experience flooding, which represents a minimum preference condition. Conversely, maximum preference is applied when higher values indicate greater flood potential, such as regions with high rainfall intensity, which is processed using the normalization shown in line 3.

Clustering using K-Means Cluster

K-Means works by partitioning data into several groups initialized with a predefined value of K (the number of clusters) [21] [22]. Each cluster contains data with similar characteristics, as similarity is measured using the Euclidean distance formula. In this study, the determination of the optimal number of clusters (K) was carried out using the Elbow, Gap Statistic, and Silhouette methods [23].

Visualization of Data Distribution

The areas belonging to each cluster were visualized in the form of histograms to observe the data distribution within each cluster. Data balance was assessed by measuring the difference in the number of members across the formed clusters. To determine whether the data distribution was balanced or not, the Coefficient of Variation (CV) was employed.

3 Results and Discussions

Data Acquisition

The study area of this research covers several regions in Central Java Province. The regions presented in Fig. 1 are identified as having a high flood potential based on data from National Disaster Management Agency (BNPB).

Figure 1. Study area

Dominance Testing using Skyline Query

The dominance test using the skyline query identified 18 regions as flood-prone, while 1,574 regions were classified as low potential. The imbalance between these two categories is highly significant. The distribution results are visualized in the histogram in Fig. 2, which clearly illustrates the frequency differences between classes.

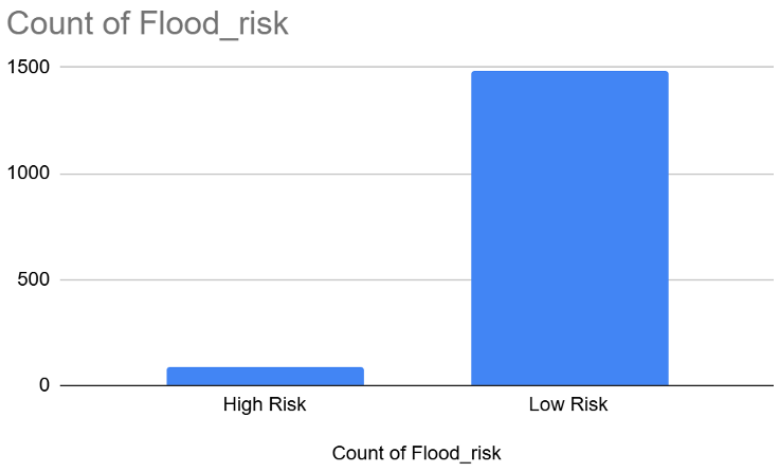
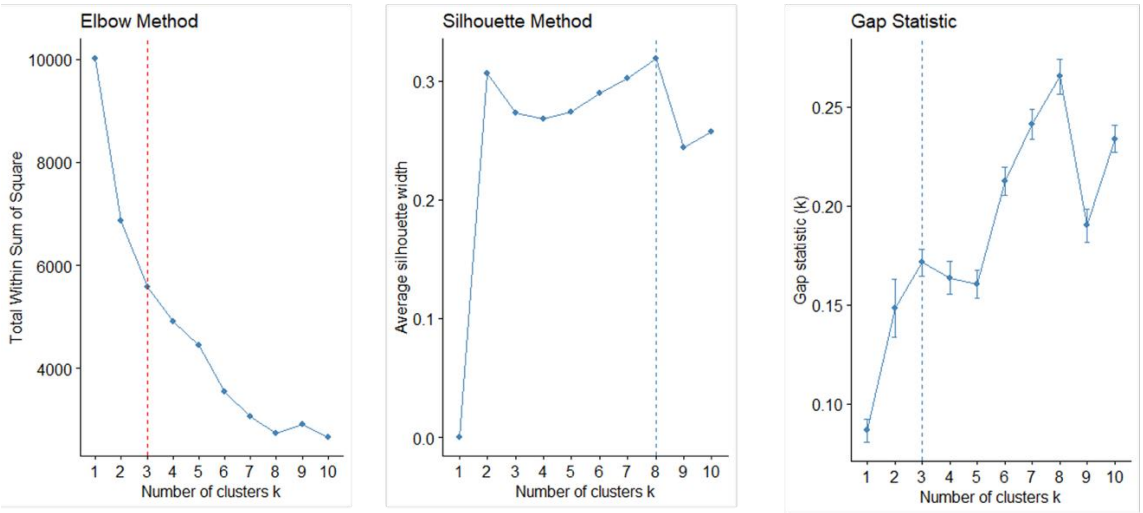


Figure 2. The results of the dominance test

Clustering using K-Means Cluster

The data clustering was carried out under three scenarios, following the approaches used to determine the most optimal value of K. The K recommendations based on the Elbow and Gap Statistic methods are shown in Fig. 1 (a)-(c). Both the Elbow and Gap Statistic methods suggested an optimal K of 3, while the Silhouette method recommended a K value of 8.

The variation in the recommended number of clusters comes from the different principles used by each evaluation method. The Elbow method, shown in Fig. 3(a), identifies K = 3 because the decrease in the sum of squared errors becomes noticeably smaller after that point, meaning additional clusters do not add much explanatory value. The Gap Statistic in Fig. 3(b) also points to K = 3, as this value gives the largest separation between the actual data clustering and the reference distribution, indicating that three clusters capture the most meaningful structure in the dataset. Meanwhile, the Silhouette method in Fig. 3(c) suggests K = 8, as this configuration provides the highest average silhouette score, reflecting clusters that are more compact and clearly separated. The Silhouette method tends to detect finer patterns in the data, which is why it favors a larger number of clusters.



(a) (b) (c)

Figure 3. The recommendation results of the methods are as follows: (a) Elbow; (b) Silhouette; (c) Gap Statistic.

Thus, choosing between K = 3 or K = 8 depends on the analysis objective: whether the study prioritizes broader, more general clustering patterns, or a more detailed clustering that highlights local differences.

Visualization of Data Distribution

The visualization of data distribution is presented in Fig. 2. Fig. 4 (a) shows the distribution when the number of clusters K=3, which corresponds to the recommendation from the Elbow and Gap Statistic methods. Meanwhile, Fig. 4 (b) illustrates the distribution resulting from K-Means clustering with K=8. Each scheme was evaluated to determine the degree of balance in cluster membership using the balance ratio and coefficient of variation (CV), as shown in Table 1.

In general, a balance ratio close to 1 and a low CV value indicate strong clustering performance because the number of members across clusters is relatively even. The results show that Scheme 1 and Scheme 3, which both generate three clusters, have a balance ratio of 1.3 and a CV of 13.9%, reflecting a relatively balanced cluster size distribution with low variation. In contrast, Scheme 2 has a balance ratio of 2.8 and a CV of 34.8%, indicating high variation and a more uneven distribution of cluster members.

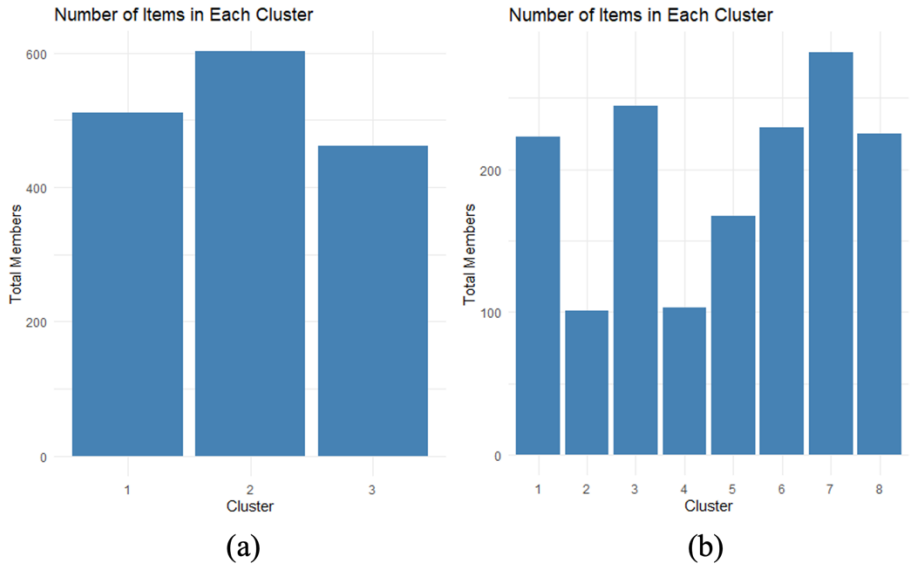
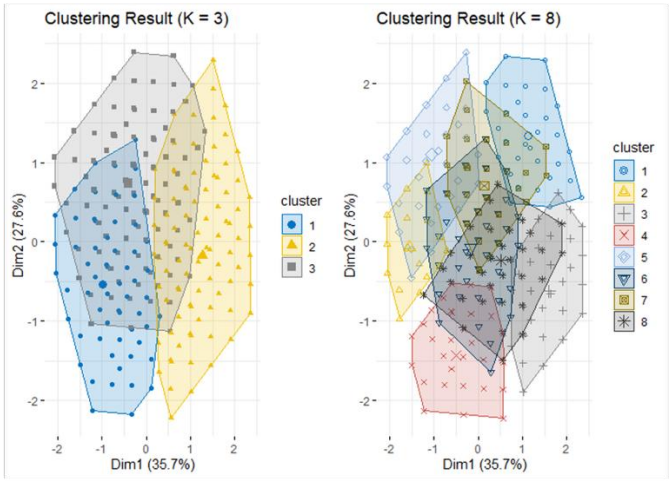


Figure 4. Cluster membership distribution: (a) k=3, (b) k=8

Therefore, Scheme 1 (Elbow) and Scheme 3 (Gap Statistic) demonstrate stronger clustering performance compared to Scheme 2 (Silhouette), which is considered less balanced. The visualization of the clustering results can be seen in Fig. 5.

Table 1. Comparison of balance ratio and coefficient of variation values

Schema	Cluster	Cluster Membership	Balance Ratio	Coefficient of Variation (%)
Schema 1, and Schema 3	C1	501	**1.3**	**13.9**
	C2	602		
	C3	461		
Schema 2	C1	223	**2.8**	**34.8**
	C2	101		
	C3	244		
	C4	103		
	C5	167		
	C6	229		
	C7	282		
	C8	225		



(a) (b)

Figure 5. Clusterring result: (a) k=3, (b) k=8

4 Conclusions

The analysis results indicate a data imbalance generated by the skyline query, with 18 areas identified as flood-prone and 1,574 areas categorized as low-risk. In determining the number of clusters, the Elbow and Gap Statistic methods recommend K=3, while the Silhouette method suggests K=8. The evaluation of cluster membership distribution shows that Schemes 1 and 3 are more balanced compared to Scheme 2. Future research should focus on the relevance of flood classes with spatial and environmental factors. For machine learning applications, it is recommended to apply imbalance handling techniques such as SMOTE or cost-sensitive learning, along with ensemble algorithms, to improve model accuracy in identifying the limited number of flood-prone areas.

Acknowledgements

The authors would like to express their gratitude to the Research and Community Service Institute of Universitas Siliwangi for providing financial support in the 2025 fiscal year, which enabled this research to be successfully conducted. Sincere thanks are also extended to all parties who contributed to the research process and the preparation of this article.

References

[1] B. I. Nasution, F. M. Saputra, R. Kurniawan, A. N. Ridwan, A. Fudholi, and B. Sumargo, “Urban vulnerability to floods investigation in jakarta, Indonesia: A hybrid optimized fuzzy spatial clustering and news media analysis approach,” *International Journal of Disaster Risk Reduction*, vol. 83, p. 103407, Dec. 2022, doi: 10.1016/j.ijdrr.2022.103407.

[2] N. T. Lapatta, “Ecotourism Recommendations based on Sentiments Using Skyline Query and Apache-Spark,” *Journal of Social Science*, vol. 3, no. 3, pp. 534–546, May 2022, doi: 10.46799/jss.v3i3.333.

[3] V. Purwayoga, S. Yuliyanti, A. Nurkholis, H. Gunawan, S. Sokid, and N. Kartini, "Distribution Model of Personal Protective Equipment (PPE) Using the Spatial Dominance Test and Decision Tree Algorithm," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 3, p. 1445, Sep. 2024, doi: 10.62527/joiv.8.3.2471.

[4] V. Purwayoga and B. Susanto, "Rekomendasi Daerah Penyalur Tenaga Kesehatan Covid-19 Dengan Menggunakan Skyline Query," *Fountain of Informatics Journal*, vol. 7, no. 1, p. 22, Oct. 2021, doi: 10.21111/fij.v7i1.5720.

[5] V. Purwayoga, "Modified skyline query to measure priority region for personal protective equipment recipient of COVID-19 health workers," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 3, pp. 167–173, Jul. 2021, doi: 10.14710/jtsiskom.2021.14003.

[6] V. Purwayoga, E. N. F. Dewi, and Z. Gusnadi, "B-Shinance: Aplikasi R-Shiny Interaktif untuk Percepatan Visualisasi Daerah Potensi Banjir Berdasarkan Uji Dominasi," *JASIEK (Jurnal Aplikasi Sains, Informasi, Elektronika dan Komputer)*, vol. 5, no. 2, pp. 69–76, Dec. 2023, doi: 10.26905/jasiek.v5i2.11546.

[7] V. Purwayoga and Z. Gusnadi, "Perbandingan Metode Pengukuran Jarak pada Analisis Potensi Banjir Menggunakan Spatial Skyline Query," *Infomatek*, vol. 27, no. 1, pp. 87–94, Jun. 2025, doi: 10.23969/infomatek.v27i1.24149.

[8] M. A. Mohamud *et al.*, "A Systematic Literature Review of Skyline Query Processing Over Data Stream," *IEEE Access*, vol. 11, pp. 72813–72835, 2023, doi: 10.1109/ACCESS.2023.3295117.

[9] R. Amin, "Development of Skyline Query Algorithm for Individual Preference Recommendation in Streaming Data," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 1012–1025, May 2025, doi: 10.47738/jads.v6i2.599.

[10] A. Vlachou, C. Doulkeridis, J. B. Rocha-Junior, and K. Nørvåg, "Decisive skyline queries for truly balancing multiple criteria," *Data Knowl Eng*, vol. 147, p. 102206, Sep. 2023, doi: 10.1016/j.datak.2023.102206.

[11] V. Purwayoga, M. Al Husaini, and H. H. Lukmana, "Visualisasi Skyline Query untuk Distribusi Tenaga Kesehatan COVID-19," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 9, no. 1, Apr. 2023, doi: 10.28932/jutisi.v9i1.5624.

[12] Ainaa Hanis Zuhairi, Fitri Yakub, Mas Omar, Muhammad Sharifuddin, Khamarrul Azahari Razak, and Amrul Faruq, "Imbalanced Flood Forecast Dataset Resampling Using SMOTE-Tomek Link," *Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES)*, vol. 10, pp. 845–850, Oct. 2024, doi: 10.5109/7323359.

[13] D. Dablain, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans Neural Netw Learn Syst*, 2022, doi: 10.1109/TNNLS.2021.3136503.

[14] J. C. Munguía Mondragón, E. Rendón Lara, R. Alejo Eleuterio, E. E. Granda Gutirrez, and F. Del Razo López, "Density-Based Clustering to Deal with Highly Imbalanced Data in Multi-Class Problems," *Mathematics*, vol. 11, no. 18, p. 4008, Sep. 2023, doi: 10.3390/math11184008.

[15] F. Reviantika, C. N. Harahap, and Y. Azhar, "Analisis Gempa Bumi Pada Pulau Jawa Menggunakan Clustering Algoritma K-Means Analysis Earthquake in Java Island Using Clustering K-Means Algorithm," *Jurnal Dinamika Informatika*, vol. 9, no. 1, 2020, [Online]. Available: https://twitter.com/infobmkg

[16] Michael Kevin Adinata, Ali Mahmudi, and Yosep Agus Pranoto, "Klasterisasi Daerah Rawan Bencana Alam Menggunakan Algoritma K-Means," *Infotek: Jurnal*

Informatika dan Teknologi, vol. 8, no. 1, pp. 250–260, Jan. 2025, doi: 10.29408/jit.v8i1.28196.

[17] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," *Journal of Applied Informatics and Computing*, vol. 7, no. 1, pp. 89–97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.

[18] D. Hartama, W. Wanayumini, and I. S. Damanik, "Pengelompokan Algoritma K-Means dan K-Medoid Berdasarkan Lokasi Daerah Rawan Bencana di Indonesia dengan Optimasi Elbow, DBI, dan Silhouette," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, Sep. 2024, doi: 10.47065/bits.v6i2.5851.

[19] A. Parven *et al.*, "Impacts of disaster and land-use change on food security and adaptation: Evidence from the delta community in Bangladesh," *International Journal of Disaster Risk Reduction*, vol. 78, p. 103119, Aug. 2022, doi: 10.1016/j.ijdrr.2022.103119.

[20] A. Annisa and S. Khairina, "Location Selection Based on Surrounding Facilities in Google Maps using Sort Filter Skyline Algorithm," *Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika*, vol. 7, no. 2, pp. 65–72, Jul. 2021, doi: 10.23917/khif.v7i2.12939.

[21] V. Purwayoga and B. Susanto, "Pengelompokkan Daerah Berdasarkan Ketersediaan Masjid Muhammadiyah Dengan Algoritma K-Means," *Jurnal Teknologi,* vol. 13, no. 1, 2021, doi: 10.24853/jurtek.13.1.75-80.

[22] N. H. Wulandari and V. Purwayoga, "Cluster Change Analysis to Assess the Effectiveness of Speaking Skill Techniques using Machine Learning,"

International Journal of Applied Sciences and Smart Technologies, vol. 7, no. 1, pp. 1–14, Jun. 2025, doi: 10.24071/ijasst.v7i1.9667.

[23] V. Purwayoga, "Optimasi Jumlah Cluster pada Algoritme K-Means untuk Evaluasi Kinerja Dosen," *Jurnal Informatika Universitas Pamulang*, vol. 6, no. 1, p. 118, Mar. 2021, doi: 10.32493/informatika.v6i1.9522.